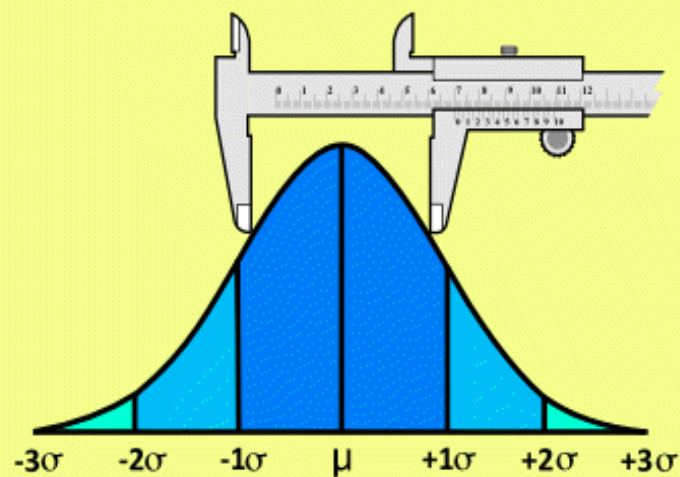
KARL GRUBE **GERHARD GRUBE**

AMPLITUDE
PONTOS DISCREPANTES
TAMANHO DE AMOSTRA
EM ESTATÍSTICA

Curitiba, Junho 2012

AMPLITUDE, PONTOS DISCREPANTES E TAMANHO DE AMOSTRA EM ESTATÍSTICA
Por Karl Grube e Gerhard Grube, Junho de 2012. Revisão 26/01/2014

APLICAÇÃO DOS MÉTODOS DESENVOLVIDOS NESTE TRABALHO

Para facilitar a aplicação dos métodos desenvolvidos neste trabalho, apresentamos a seguir um resumo de como utilizar as tabelas. As justificativas e maiores detalhes encontram-se no corpo do trabalho. Ao usar as tabelas, valores intermediários podem ser obtidos por interpolação. O objeto é sempre a análise de amostras aleatórias de uma variável contínua com distribuição aproximadamente normal.

TABELAS 1 e 2 – DESVIO MÁXIMO E AMPLITUDE DE AMOSTRAS NORMAIS

Estas tabelas refletem o fato que, quanto maior tamanho da amostra, maior é a probabilidade de se encontrar pontos muito afastados da média. Tanto a amplitude, como o desvio máximo aumentam.

Amplitude é a diferença entre os dois valores extremos da amostra: $x_{max} - x_{min}$.

Desvio máximo é o afastamento do ponto mais distante da média: $x_{max} - \mu$.

A tabela 1 apresenta os desvios máximos, em função do tamanho N da amostra. A tabela 2 apresenta as amplitudes (metade). Os valores de Z indicados nas tabelas não são exatos; são os que, em média, podem ser esperados. As respectivas equações estão indicadas no final das tabelas.

TABELA 1 – DESVIO MÁXIMO

N	Z	N	Z	N	Z
2	0,674	20	2,099	200	2,922
3	1,052	30	2,263	300	3,046
4	1,264	40	2,374	400	3,132
5	1,408	50	2,456	500	3,197
6	1,516	60	2,522	600	3,249
7	1,602	70	2,576	700	3,293
8	1,673	80	2,622	800	3,331
9	1,733	90	2,662	900	3,363
10	1,786	100	2,698	1000	3,392

N: número de pontos da amostra

$Z = (x_{max} - \mu) / \sigma$

(μ é a média, σ é o desvio padrão)

TABELA 2 – AMPLITUDE

N	Z	N	Z	N	Z
2	0,431	20	1,909	200	2,776
3	0,802	30	2,084	300	2,905
4	1,022	40	2,201	400	2,994
5	1,175	50	2,287	500	3,062
6	1,289	60	2,356	600	3,116
7	1,381	70	2,414	700	3,161
8	1,457	80	2,462	800	3,199
9	1,521	90	2,504	900	3,234
10	1,576	100	2,542	1000	3,264

N: número de pontos da amostra

$Z = (x_{max} - x_{min}) \times 0,5 / \sigma$

(σ é o desvio padrão)

Os valores de Z dependem da distribuição normal. A interpolação pode ser evitada usando-se uma planilha (Open Office, Excel 2010); os valores de Z podem ser obtidos em função de N pelas expressões:

Para o desvio máximo: =INV.NORM(0,5^(1/(B13-1))/2 +0,5;0;1)

Para a amplitude: =INV.NORM((1/3)^(1/(B13-1))/2 +0,5;0;1)

nestas expressões, B13 é a célula que contém N (número de pontos da amostra).

EXEMPLOS:

a) Numa amostra de 5 pontos, o valor máximo é 15 e o mínimo é 7. Quais são os valores aproximados da média e do desvio padrão?

A média aproximada é (15 + 7) / 2 = 11

A metade da amplitude é (15 - 7) / 2 = 4. Da tabela 2 acima, para N = 5, obtém-se Z = 1,175. O desvio padrão aproximado é: $\sigma = 4 / 1,175 = 3,4$

Observação: O intervalo de variação da média pode ser estimado utilizando os desvios máximos da tabela 1 (para 5 pontos, z = 1,408):

Valor máximo da média: 7 + 1,408 × 3,4 = 11,8

Valor mínimo da média: 15 - 1,408 × 3,4 = 10,2

b) Numa amostra de 3 pontos, cuja média é 15, o maior valor é 22. Qual é o maior valor que pode ser esperado numa amostra de 200 pontos?

Da tabela 1 obtemos:

para N = 3, Z = 1,052

para N = 200, Z = 2,922

$Z = (x_{max} - \mu) / \sigma$ (o valor de σ não precisa ser calculado)

$x_{max} - \mu = 2,922 / 1,052 \times (22 - 15) = 19,4$

O valor máximo é: $x_{max} = 15 + 19,4 = 34,4$

TABELA 3 – IDENTIFICAÇÃO DE PONTOS DISCREPANTES

A identificação de um ponto estatisticamente discrepante é importante porque este ponto é uma indicação de que ocorreu uma anomalia no processo, um erro na medição, um erro grosseiro, ou uma flutuação estatística excepcional. Pontos estatisticamente discrepantes podem distorcer a estimativa da média. Por este motivo, é interessante eliminá-los da amostra.

Para identificar os pontos discrepantes, usa-se a tabela 3. Dada uma amostra, calcula-se a média e o desvio padrão. A seguir, divide-se a diferença entre o valor de um ponto e a média, pelo desvio padrão. Compara-se o resultado z, em valor absoluto, com o limite z_d da tabela. Se for maior, o ponto é discrepante:

$z = |(x - \bar{x}) / \sigma|$

$\bar{x}$: média da amostra

σ: desvio padrão da amostra

x é discrepante se $z > z_d$

Os pontos da amostra devem estar ordenados. Examina-se, inicialmente, o ponto mais afastado da média. Se ele for discrepante, é eliminado, observando-se o seguinte.

O ponto discrepante não deve ser eliminado, se a diferença em relação à média for menor que a diferença admissível *d*. A diferença admissível *d* é a maior diferença que ainda não é considerada significativa. É o erro aceitável. Ao definir um valor, deve-se ter em mente que erros menores que 1% são difíceis de serem obtidos. Em pesquisa tecnológica, às vezes se aceita erros de 10% ou mais.

Eliminado o ponto, recalcula-se a média e o desvio padrão. O processo deve ser repetido, até que todos os pontos discrepantes sejam eliminados, ou até que tenham sido eliminados 1/3 dos pontos (a amostra restante não deverá conter menos de 2/3 dos pontos originais). A média e o desvio padrão, calculados com a amostra remanescente, representam melhor os valores verdadeiros.

TABELA 3 – LIMITES DISCREPANTES – MÉTODO PROPOSTO

N	Zd	N	Zd	N	Zd
2	---	20	2,231	200	3,025
3	1,121	30	2,388	300	3,145
4	1,391	40	2,494	400	3,229
5	1,565	50	2,573	500	3,292
6	1,672	60	2,637	600	3,343
7	1,754	70	2,691	700	3,386
8	1,822	80	2,734	800	3,422
9	1,881	90	2,773	900	3,454
10	1,931	100	2,807	1000	3,483

Nesta tabela, N é o número de pontos da amostra e Zd é o limite discrepante.

Para amostras de 6 ou mais pontos, pode-se evitar a interpolação, calculando os limites discrepantes z_d numa planilha (Open Office, Excel 2010), através da expressão:

=INV.NORM(0,608914^(1/(D13 -1))/2 +0,5;0;1)

onde D13 é a célula que contém N (número de pontos da amostra).

EXEMPLO:
Pretende-se adquirir um eletrodoméstico com preço aproximado de R$ 1000,00. Neste nível de custo, considera-se que uma diferença de R$ 20,00 não é significativa (*d* = 20). Foram obtidos 3 preços:

a) R$ 800,00, R$ 1000,00 e R$ 1700,00. O preço mais alto é discrepante?
Aplicando o método proposto, temos:
média = 1166,67
desvio padrão = 472,58
diferença = 1700,00 - 1166,67 = 533,33
z = 533,33 / 472,58 = 1,128
da tabela 3, para N = 3, z_d = 1,121
como z é maior que z_d, o ponto é discrepante. A diferença é maior que *d*, portanto significativa, confirmando que o ponto deve ser eliminado.

Observação: O valor de z deve ser calculado, com precisão, até a terceira casa decimal.

b) R$ 975,00, R$ 1000,00, R$ 1000,00. O preço mais baixo é discrepante?
Aplicando o método proposto, temos:
média = 991,67
desvio padrão = 14,43
diferença = 991,67 - 975,00 = 16,67
z = 16,67 / 14,43 = 1,155
da tabela 3, para N = 3, z_d = 1,121
como z é maior que z_d, o ponto é discrepante. Porém a diferença é menor que diferença admissível *d*. O ponto não deve ser eliminado.

Observação: Quando uma amostra tem vários valores iguais, qualquer ponto um pouco diferente tenderá a ser indicado como discrepante; neste caso, o critério decisivo é a diferença admissível *d*.

TABELA 4 – TAMANHO DA AMOSTRA

Em qualquer experimento, enfrenta-se o problema de determinar o tamanho da amostra. Quanto maior a amostra, mais preciso será o resultado do experimento. Por outro lado, o custo de obtenção da amostra aumenta.
Para aplicar o método proposto, são necessários valores estimados da diferença admissível *d* e o desvio padrão σ. Com a relação *d*/σ, o tamanho N é determinado usando a tabela 4. Os tamanhos indicados são os mínimos recomendados. Nada impede que sejam usadas amostras maiores, por exemplo, quando a população amostrada é heterogênea, visando garantir que a amostra seja representativa.
Após realizar a amostragem, deve ser verificada a existência de pontos discrepantes, conforme o tópico anterior. Os pontos discrepantes devem ser eliminados e substituídos por outros, completando o tamanho N requerido.

TABELA 4 – TAMANHO DA AMOSTRA – MÉTODO PROPOSTO

N	*d*/σ	N	*d*/σ	N	*d*/σ	N	*d*/σ
1	≥1,29	12	0,201	35	0,074	90	0,031
2	0,862	14	0,174	40	0,065	100	0,028
3	0,647	16	0,154	45	0,059	110	0,026
4	0,518	18	0,138	50	0,053	120	0,024
5	0,432	20	0,125	55	0,049	130	0,022
6	0,371	22	0,115	60	0,045	140	0,021
7	0,325	24	0,106	65	0,042	150	0,020
8	0,289	26	0,098	70	0,039	160	0,019
9	0,260	28	0,091	75	0,036	180	0,017
10	0,237	30	0,086	80	0,034	200	0,015

N: número de pontos da amostra
d/σ: diferença admissível / desvio padrão estimado

Numa planilha (Open Office, Excel 2010), o valor de *d*/σ pode ser obtido mediante a expressão: =(2,58/(D13+1))*(1+ 0,001*D13)
onde D13 é a célula que contém N (número de pontos da amostra).

Em geral é possível obter uma estimativa razoável do desvio padrão da população, analisando o comportamento esperado da variável. Tendo-se uma idéia da faixa de variação, pode-se admitir, grosso modo, que ela equivale a 5 ou 6 desvios padrão.
Após realizar alguns testes (pelo menos três), o valor do desvio padrão poderá ser recalculado, revisando-se, se necessário, o tamanho da amostra.
A diferença admissível *d* é a maior diferença, na média calculada, que ainda não é considerada significativa. É o erro aceitável. Ao definir um valor, deve-se ter em mente que erros menores que 1% são difíceis de serem obtidos. Em pesquisa tecnológica, às vezes se aceita erros de 10% ou mais. De qualquer modo, uma vez definida a diferença admissível, o sistema de medição deve ser escolhido com a precisão adequada. O erro de medição deve ser bem menor que a diferença admissível.

EXEMPLO:
Deseja-se pesquisar o preço de um equipamento industrial cujo valor, numa primeira estimativa, é de R$ 150.000,00, com uma faixa de variação entre R$ 100.000,00 e R$ 200.000,00.
Quantas propostas deverão ser solicitadas? Em geral, neste tipo de pesquisa, os custos não são desprezíveis; a elaboração e a análise das propostas sempre exigem um tempo considerável. Portanto, deve-se procurar a quantidade mínima necessária.
O desvio padrão aproximado é (200.000 - 100.000) / 5 = 20.000
A diferença admissível *d* deve ser definida conforme o objetivo da pesquisa. Se, por exemplo, o objetivo for uma estimativa preliminar de custos, uma diferença de 10% do preço esperado é aceitável; portanto,
d = 150.000 × 0,10 = 15.000
d/σ = 15.000 / 20.000 = 0,75
Da Tabela 4, obtemos N = 3
Assim, para uma estimativa preliminar de custos, bastam três propostas.

AMPLITUDE, PONTOS DISCREPANTES E TAMANHO DE AMOSTRA EM ESTATÍSTICA

Karl Grube, Engenheiro Químico, formado pela UFPR
Gerhard Grube, Engenheiro Mecânico, formado pela UFPR
Curitiba, Junho de 2012.

No presente trabalho são desenvolvidos métodos para a determinação da amplitude (intervalo, "range") de uma amostra, identificação de pontos discrepantes ("outliers") e determinação do tamanho da amostra ("sample size").
Estes assuntos são pouco ventilados em livros-texto de estatística e as soluções oferecidas nem sempre são satisfatórias, quando aplicadas a problemas de engenharia. Apresentamos algumas alternativas, mais baseadas no bom senso do que em conhecimento teórico.
O trabalho está dividido em três partes, cada uma relativa a um tema. O objeto é sempre a análise de amostras aleatórias de uma variável contínua com distribuição aproximadamente normal.

1 – PRIMEIRA PARTE: AMPLITUDE DE AMOSTRAS NORMAIS

Por Karl Grube e Gerhard Grube, Junho de 2012

1.1 – INTRODUÇÃO E RESUMO

Definimos como amplitude, também chamada intervalo ("range"), de uma amostra a diferença entre os dois valores extremos da amostra. Desvio máximo é o afastamento do ponto mais distante da média.
É sabido que a amplitude aumenta com o tamanho da amostra. Quanto maior a amostra, maior a probabilidade de se encontrar valores muito altos ou muito baixos.
O conhecimento da relação entre o número de pontos da amostra e a amplitude é útil em diversas situações:
–quando se quer saber quais valores poderão ser atingidos em uma amostra de determinado tamanho
–quando se deseja estimar a média e o desvio padrão conhecendo-se apenas os valores extremos de uma amostra
–quando se quer saber se um determinado ponto da amostra é compatível com o tamanho da mesma (análise de pontos discrepantes).
Nesta primeira parte procura-se desenvolver um método para determinar a amplitude e o desvio máximo de amostras normais, conforme definidos acima. Analisando as probabilidades da distribuição normal, são deduzidas equações para a amplitude e para o desvio máximo. Os resultados numéricos estão apresentados no Item 1.3.
As equações foram verificadas experimentalmente por meio de algumas amostras normais (anexo 1) e de um grande número de amostras aleatórias (anexo 2). Aproximadamente, os resultados confirmam as equações teóricas.

1.2 – DESENVOLVIMENTO TEÓRICO DO MÉTODO

A amplitude de uma amostra normal está relacionada com as probabilidades da distribuição normal. Dado um intervalo ± z σ, a probabilidade de um ponto qualquer cair neste intervalo é p e a probabilidade do ponto cair fora é 1 - p. A probabilidade de um segundo ponto cair no intervalo é p^2, enquanto a probabilidade do segundo ponto cair fora é 1 - p^2.
E assim por diante. A probabilidade p pode ser obtida de tabelas da distribuição normal em função do valor de z. Consideremos um exemplo. Dado um intervalo ± 1,5 σ, obtém-se da tabela p = 0,866, que é a probabilidade de um ponto qualquer cair neste intervalo. A probabilidade dele cair fora é 1 - 0,866 = 0,134. A probabilidade de um segundo ponto cair no intervalo é 0,866^2 = 0,750,a probabilidade dele cair fora é 1 - 0,750 = 0,250.
À medida que aumenta o número de pontos no intervalo, diminui a probabilidade do próximo ponto cair dentro, enquanto a probabilidade de cair fora aumenta. Com cinco pontos no intervalo, a probabilidade do próximo ponto cair fora já é 0,513, ou seja, maior do que cair dentro.
Quando a probabilidade do último ponto da amostra cair dentro do intervalo for igual à de cair fora, admite-se que o ponto está localizado exatamente sobre o limite do intervalo. Isto permite determinar a posição do último ponto (figuras A e B).

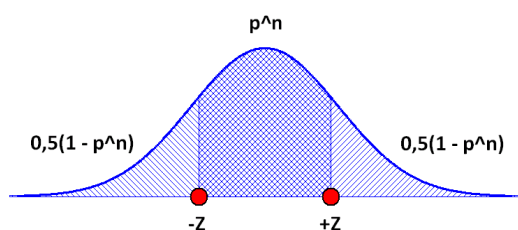
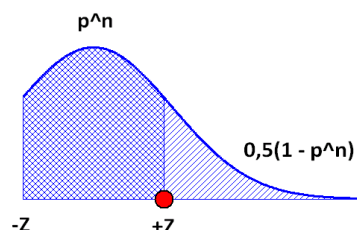



FIGURA A FIGURA B

a) Desvio máximo
Considerando o exposto acima, a posição do ponto mais afastado da média, que pode estar localizado à direita ou à esquerda da média (figura A), pode ser determinada fazendo:

p^n = 1 - p^n
2 p^n = 1
p^n = 1/2
donde se obtém

p = (1/2)^(1/n) equação [1]

n é o número de pontos no intervalo (o número de pontos na amostra é N = n+1).
Com o valor de n calcula-se p. Da tabela de distribuição normal, obtém-se o afastamento z. Vejamos, por exemplo, uma amostra de três pontos. Neste caso, n = 2, portanto p = 0,7071. Da tabela, obtém-se z = 1,052. Este é o desvio máximo esperado.

b) Amplitude

A amplitude pode ser obtida com raciocínio semelhante, porém considerando que o ponto está somente num dos lados da curva (figura B). Neste caso,

$p^n = (1 - p^n) / 2$

donde se obtém

$p = (1/3)^{(1/n)}$ equação [2]

Por exemplo, numa amostra de três pontos, n = 2, portanto p = 0,5774. Da tabela da distribuição normal, obtemos z = 0,802. A amplitude esperada é 2 × 0,802 = 1,604.

c) Limite discrepante

Se a probabilidade do último ponto da amostra cair fora do intervalo for menor do que cair dentro, este tende a ser discrepante (isto é, um ponto não pertencente à população). Genericamente o limite pode ser definido por

$p^n = g\ (1 - p^n)$ equação [3]

onde g é um fator maior que 1. Este fator é, até certo ponto, arbitrário. Quanto maior o seu valor, menos pontos serão considerados discrepantes. Mais adiante, na segunda parte deste trabalho, esta concepção é usada para elaborar um método de identificação de pontos discrepantes.

1.3 – AMPLITUDE E DESVIO MÁXIMO EM AMOSTRAS NORMAIS

A tabela 1 abaixo apresenta os desvios máximos calculados coma equação teórica, em função do tamanho N da amostra. A tabela 2 apresenta as amplitudes (metade).

Os valores de Z são os médios esperados; não são valores exatos. Valores intermediários podem ser obtidos por interpolação. As respectivas equações encontram-se no final das tabelas.

TABELA 1 – DESVIO MÁXIMO

N	Z	N	Z	N	Z
2	0,674	20	2,099	200	2,922
3	1,052	30	2,263	300	3,046
4	1,264	40	2,374	400	3,132
5	1,408	50	2,456	500	3,197
6	1,516	60	2,522	600	3,249
7	1,602	70	2,576	700	3,293
8	1,673	80	2,622	800	3,331
9	1,733	90	2,662	900	3,363
10	1,786	100	2,698	1000	3,392

N: número de pontos da amostra

$Z = (x_{max} - \mu) / \sigma$ (x_{max} é o maior ponto da amostra, μ é a média e σ é o desvio padrão)

TABELA 2 – AMPLITUDE

N	Z	N	Z	N	Z
2	0,431	20	1,909	200	2,776
3	0,802	30	2,084	300	2,905
4	1,022	40	2,201	400	2,994
5	1,175	50	2,287	500	3,062
6	1,289	60	2,356	600	3,116
7	1,381	70	2,414	700	3,161
8	1,457	80	2,462	800	3,199
9	1,521	90	2,504	900	3,234
10	1,576	100	2,542	1000	3,264

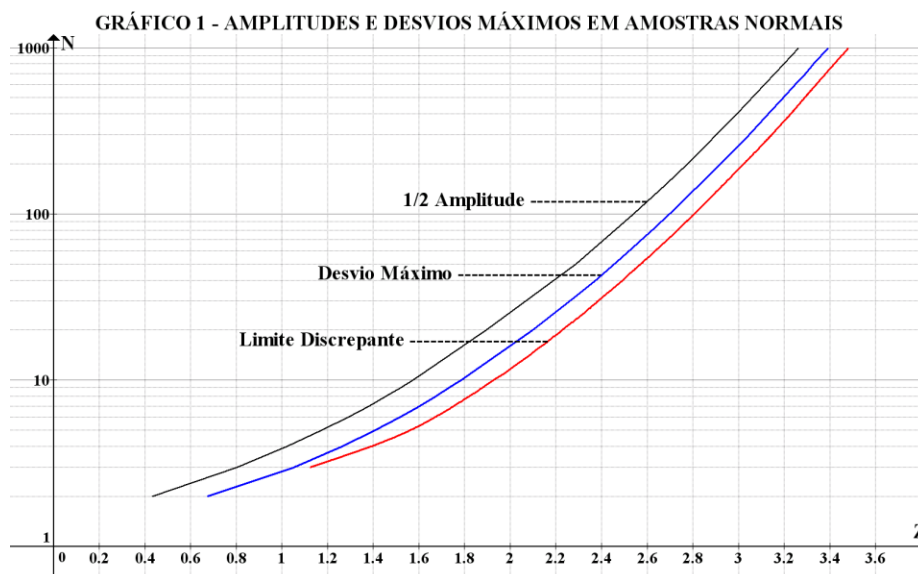
N: número de pontos da amostra, $Z = (x_{max} - x_{min}) \times 0,5 / \sigma$
(x_{max} e x_{min} são, respectivamente, o maior e o menor ponto da amostra; σ é o desvio padrão).

Numa planilha (Open Office, Excel 2010) os valores de Z podem ser obtidos em função de N pelas expressões:

Para o desvio máximo: =INV.NORM(0,5^(1/(B13-1))/2 +0,5;0;1)

Para a amplitude: =INV.NORM((1/3)^(1/(B13-1))/2 +0,5;0;1)

em que B13 é a célula que contém N (número de pontos da amostra).

Verifica-se que a amplitude aumenta indefinidamente com o tamanho da amostra. Quanto maior a amostra, maior será o afastamento de um ponto extremo.
Em amostras de 3 pontos pode-se esperar que o ponto mais afastado (desvio máximo) esteja a cerca de 1σ da média; em amostras de 17 pontos o afastamento máximo é de 2σ; com 260 pontos, chega a 3σ. O gráfico 1 a seguir apresenta as curvas da amplitude e do desvio máximo, conforme as tabelas acima. Também está desenhada a curva dos limites discrepantes, obtidos pelo método desenvolvido na segunda parte.

N: número de pontos da amostra

Z: afastamento da média = $(x - \bar{x}) / \sigma$

1.4 – VERIFICAÇÃO DO MÉTODO

Com o objetivo de verificar a exatidão das equações teóricas desenvolvidas, foram realizadas duas verificações experimentais. Numa verificação mais grosseira (anexo 1), foram utilizadas algumas amostras hipotéticas aproximadamente normais. Para uma verificação experimental mais exaustiva (anexo 2) foram obtidas as estimativas do desvio máximo e da amplitude para dez mil amostras extraídas aleatoriamente da distribuição normal. Os detalhes constam dos anexos. Aproximadamente, os resultados experimentais confirmam as equações teóricas.

1.5 – EXEMPLOS

a) Numa amostra de 5 pontos, o valor máximo é 15 e o mínimo é 7. Quais são os valores aproximados da média e do desvio padrão?
A média aproximada é (15 + 7) / 2 = 11
A metade da amplitude é (15 - 7) / 2 = 4. Da tabela 2 acima, para N = 5, obtém-se Z = 1,175. O desvio padrão aproximado é σ = 4 / 1,175 = 3,4

Observação: O intervalo de variação da média pode ser estimado utilizando os desvios máximos da tabela 1 (para 5 pontos, z = 1,408):
Valor máximo da média: 7 + 1,408 × 3,4 = 11,8
Valor mínimo da média: 15 - 1,408 × 3,4 = 10,2

b) Numa amostra de 3 pontos, cuja média é 15, o maior valor encontrado é 22. Qual é o maior valor que pode ser esperado numa amostra de 200 pontos?
Da tabela 1 obtemos:
para N = 3, Z = 1,052
para N = 200, Z = 2,922
$Z = (x_{max} - \mu) / \sigma$
O valor de σ não precisa ser calculado:
$x_{max} - \mu = 2{,}922 / 1{,}052 \times (22 - 15) = 19{,}4$
O valor máximo é $x_{max} = 15 + 19{,}4 = 34{,}4$

ANEXOS RELATIVOS À PRIMEIRA PARTE

ANEXO 1: Verificação com algumas amostras normais.
ANEXO 2: Verificação com amostras aleatórias

2 – SEGUNDA PARTE: IDENTIFICAÇÃO DE PONTOS DISCREPANTES

Por Karl Grube e Gerhard Grube, Junho de 2012

2.1 – INTRODUÇÃO E RESUMO

Estatisticamente, ponto discrepante ("outlier") é um ponto que está muito afastado da média de uma amostra, sendo improvável que pertença à população.
A identificação de um ponto estatisticamente discrepante é importante, porque este ponto sempre é uma indicação de que ocorreu:
–uma anomalia no processo
–um erro na medição
–um erro grosseiro
–uma flutuação estatística excepcional
Um ponto estatisticamente discrepante pode distorcer a estimativa da média. Por este motivo, é interessante eliminá-lo da amostra.
Nesta segunda parte, analisa-se a aplicação do método do teste t de Student, recomendado na literatura. Esse método, no caso de amostras pequenas, pode indicar limites superiores altos demais (aceitando pontos que, pelo bom senso, deveriam ser rejeitados) e limites inferiores muito baixos (aceitando valores negativos, que não têm sentido em muitas situações reais). Em amostras maiores, tende a rejeitar pontos válidos pertencentes a amostras normais, o que também é um contra-senso.
Propõe-se um método que procura evitar estas deficiências utilizando dois critérios.
O critério 1, para amostras de até 6 pontos, parte de uma hipótese inicial diversa da adotada no método do teste t, resultando em limites mais adequados, principalmente para grandezas que são, por natureza, não-negativas.
Para amostras maiores, foi elaborado o critério 2, que leva em conta o aumento da amplitude com o tamanho da amostra, evitando que pontos normais sejam rejeitados.
Os dois critérios estão reunidos em uma tabela de limites discrepantes em função do tamanho da amostra, apresentada no Item 2.4.
O método proposto foi comparado com outros métodos. No anexo 4 comenta-se o critério de Chauvenet, que apresenta resultados em parte semelhantes. No anexo 5 os limites são comparados com os recomendados por Grubbs. Outro método citado na literatura, o "box&whisker", é discutido no anexo 6.

2.2 – O MÉTODO DO TESTE t DE STUDENT

A identificação de pontos discrepantes consiste em determinar o limite x_d, além do qual um ponto da amostra é considerado discrepante. Na literatura (Ref. 1) recomenda-se utilizar o método do teste t de Student para duas médias amostrais.
Neste teste supõe-se, como hipótese inicial, que as duas médias são iguais. Esta hipótese é geralmente conveniente, porque o "pool" das amostras permite uma estimativa melhor dos parâmetros (Ref. 2, pág. 240). Conseqüentemente, a diferença é referida ao desvio padrão combinado das duas amostras:
$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(\text{variância da média 1} + \text{variância da média 2})}$

onde $\bar{x}_1$ e $\bar{x}_2$ são as médias das duas amostras. Ao aplicar o método para identificar pontos discrepantes, considera-se que a amostra 2 contém só um ponto (o ponto analisado). Obtém-se, para o valor discrepante:

$x_d = \bar{x} \pm t \; \sigma \sqrt{(1/n + 1)}$ equação [4]

Onde $\bar{x}$, σ e n são os valores para a amostra excluindo o ponto em análise. O valor de t é obtido de tabelas para n - 1 graus de liberdade conforme o nível de confiança desejado.
No anexo 3 estão apresentados os cálculos de x_d com os quais foram traçadas as curvas do gráfico 2 abaixo, designadas pelos níveis de significância α = 0,025 e α = 0,05.
Neste gráfico, os valores são da amostra sem o ponto em análise. A ordenada n é igual ao tamanho da amostra menos um. A abcissa indica os valores discrepantes x_d para uma média igual a zero e desvio padrão igual a um.
Verifica-se que, para amostras muito pequenas, os valores de x_d obtidos por este método são bastante altos. Diminuem à medida que o número de pontos aumenta, tendendo a ficar constantes.

2.3 – O MÉTODO PROPOSTO

a) Critério 1 – Para amostras pequenas
Ao contrário do método anterior, partimos da hipótese de que as médias das duas amostras são diferentes. Não há vantagem em incluir a amostra 2 na estimativa dos parâmetros, já que ela consiste de um só ponto, que ainda é suspeito de estar "fora". Portanto, a diferença é referida apenas ao desvio padrão da média da amostra 1 (ou seja, da amostra excluindo o ponto em análise). O valor discrepante passa a ser:

$x_d = \bar{x} \pm k \, \sigma / \sqrt{n}$ equação [5]

Consideramos o fator k constante, igual para todos os tamanhos de amostra. Deste modo, o cálculo de x_d fica extremamente simples, sem necessidade de recorrer a tabelas.
Para utilizar a fórmula, é preciso estabelecer um valor adequado para k. Este valor será determinado com base em duas premissas:
–Inúmeras grandezas reais, tais como massa, volume, energia, produção, preços, etc. são, por natureza, positivas. Embora apresentem, freqüentemente, distribuições normais, não podem assumir valores negativos.
–Pode-se considerar que praticamente todos os pontos de uma população normal estão compreendidos no intervalo ± 3σ.
Em conseqüência impõe-se, como limite discrepante inferior, o valor zero. Para que a probabilidade de aceitar valores negativos seja pequena, o limite discrepante inferior deverá estar a -3σ da média.
Uma amostra de três pontos, nestas condições, poderia ser a seguinte:
$x_1 = 2 \;\; x_2 = 3 \;\; x_3 = 4$
para a amostra completa (3 pontos) obtemos $\bar{x} = 3$ e σ = 1,
o limite discrepante inferior é $x_d = \bar{x} - 3\,\sigma = 3 - 3 \times 1 = 0$

O valor de k pode ser obtido transformando a fórmula [5] acima e fazendo $x_d = 0$, com os demais valores calculados sem o primeiro ponto ($\bar{x} = 3,5$, $\sigma = 0,7071$, $n = 2$).
Obtém-se k = 7,0 e a fórmula para o limite discrepante fica

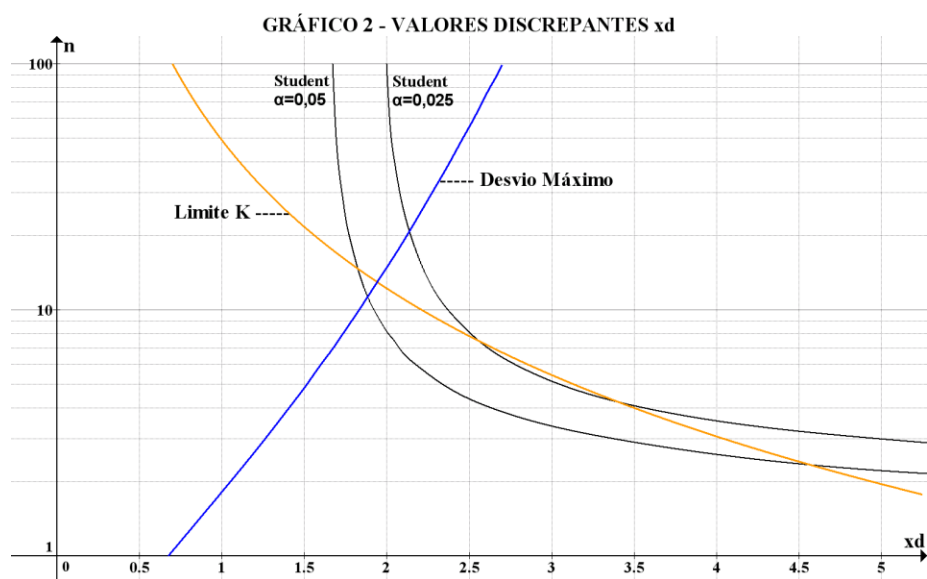
$$x_d = \bar{x} \pm 7\, \sigma / \sqrt{n} \qquad \text{equação [6]}$$

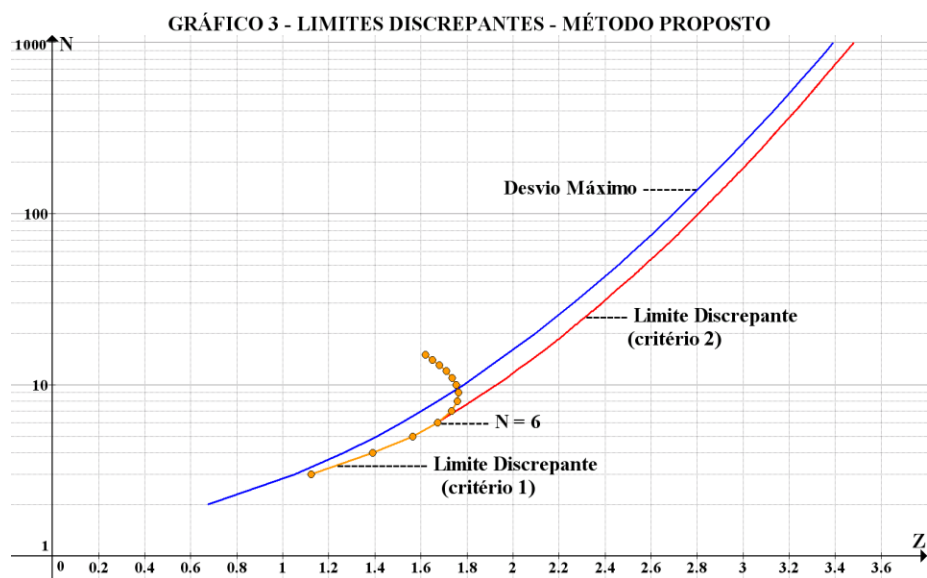
Com os limites determinados pela equação [6], praticamente todos os pontos negativos serão rejeitados, se o coeficiente de dispersão (σ/μ) for menor que 1/3. Com coeficientes de dispersão maiores, a probabilidade de aceitar valores negativos aumenta. Assim, as grandezas que podem assumir valores negativos também estão consideradas. A equação [6] tem aplicação geral.
Esta é a forma mais adequada para tratar o problema. O recurso da transformação log-normal, para evitar os valores negativos, nem sempre é correto e, dependendo dos limites, a solução não é satisfatória (ver o anexo 5, item c).
Com a equação [6] acima foi traçada a curva designada por Limite K no gráfico 2 abaixo. Os valores de x_d diminuem quando n aumenta, tendendo a se aproximar da média. Verifica-se que os valores de x_d, para amostras muito pequenas, são bem menores que os do método do teste t.

b) Critério 2 – Para amostras maiores
Um fato importante, não considerado no método do teste t, é o seguinte. A amplitude de uma amostra normal aumenta com o tamanho da amostra. Por outro lado os valores de xd definidos acima diminuem com o tamanho da amostra.
Para que os pontos de uma amostra normal não sejam considerados discrepantes, o limite deve ficar sempre à direita da curva dos desvios máximos, definida na primeira parte. Isto significa que, nas amostras maiores, o critério deve mudar.
A situação fica mais clara no gráfico 3, que considera a amostra completa, incluindo o ponto em análise (a ordenada agora é N, o tamanho da amostra).
A curva à esquerda representa o desvio máximo de amostras normais. As bolinhas amarelas representam amostras normais, nas quais o último ponto foi substituído pelo valor discrepante xd, calculado pelo critério 1.
Começando à direita da curva do desvio máximo, os pontos x_d (bolinhas amarelas) inicialmente se afastam, depois se aproximam novamente da curva, cruzando-a em aproximadamente N=9. A partir daí (pelo critério 1), pontos normais seriam considerados discrepantes. Para evitar isto, o limite discrepante (pelo critério 2) deve prosseguir paralelamente e à direita da curva do desvio máximo. Um início adequado é o ponto que está mais afastado desta curva, N=6. A partir deste ponto, o limite discrepante é determinado pela equação [3] apresentada na primeira parte:
p^n = g (1 - p^n)
Para N = 6, o limite discrepante pelo critério 1 é $z_d = 1,6723$. Da curva normal, a probabilidade de um ponto estar no intervalo ± 1,6723 σ é 0,90554. Com n = 5, encontra-se g = 1,5568. Resolvendo para o valor de p, obtemos:

$$p = 0,6089^{(1/n)} \qquad \text{equação [7]}$$

Com estes valores de p, os limites pelo critério 2 são obtidos da distribuição normal. A curva segundo o critério 2 (em vermelho, no gráfico 3) apresenta, coincidentemente,

valores muito próximos dos obtidos por Chauvenet (Ref. 3). O critério de Chauvenet é comentado no anexo 4.

n: número de pontos da amostra menos um

xd: valor discrepante para média $\bar{x}$ = 0 e desvio padrão σ = 1

N: número de pontos da amostra

Z: afastamento da média = (x - $\bar{x}$) / σ

2.4 – LIMITES DISCREPANTES – MÉTODO PROPOSTO

O método proposto pode ser aplicado mediante o uso de uma única tabela englobando os dois critérios:

TABELA 3 – LIMITES DISCREPANTES – MÉTODO PROPOSTO

N	Zd	N	Zd	N	Zd
2	----	20	2,231	200	3,025
3	1,121	30	2,388	300	3,145
4	1,391	40	2,494	400	3,229
5	1,565	50	2,573	500	3,292
6	1,672	60	2,637	600	3,343
7	1,754	70	2,691	700	3,386
8	1,822	80	2,734	800	3,422
9	1,881	90	2,773	900	3,454
10	1,931	100	2,807	1000	3,483

Nesta tabela, N é o número de pontos da amostra e Zd é o limite discrepante. Os valores para N até 6 foram obtidos pelo primeiro critério. Os valores para amostras maiores foram obtidos pelo segundo critério. Os exemplos abaixo esclarecem como foram obtidos os valores da tabela 3.

CRITÉRIO 1 (para N ≤ 6)
Parte-se de uma amostra aproximadamente normal, de tamanho N, na qual o último ponto é substituído pelo valor discrepante x_d.
Admitindo uma amostra com N = 3: $x_1 = 8$, $x_2 = 10$, $x_3 = x_d$
Com os dois primeiros pontos, obtemos
$\bar{x}(n) = 9$, $\sigma(n) = 1{,}414$
o valor de x_d é obtido com a equação [6]:
$x_d = \bar{x} + 7\sigma / \sqrt{n} = 16{,}0$
com o último ponto discrepante, a amostra fica: $x_1 = 8$, $x_2 = 10$, $x_3 = 16$
$\bar{x}(N) = 11{,}333$, $\sigma(N) = 4{,}163$
$z_d = (16{,}0 - 11{,}333) / 4{,}163 = 1{,}121$

CRITÉRIO 2 (para N > 6)
O limite discrepante é determinado em função da probabilidade p, equação [7].
Para uma amostra com N = 9 (n = 8):
p= 0,6089^(1/n) = 0,9399
obtém-se, da curva normal, $z_d = 1{,}881$

Para amostras de 6 ou mais pontos (critério 2) os limites discrepantes z_d podem ser obtidos por meio de uma planilha (Open Office, Excel 2010) usando a expressão:

=INV.NORM(0,608914^(1/(D13-1))/2 +0,5;0;1)

em que D13 é a célula que contém N (número de pontos da amostra).

2.5 – APLICAÇÃO DO MÉTODO PROPOSTO

O uso da tabela 3 é simples. Dada uma amostra, calcula-se a média e o desvio padrão. A seguir, divide-se a diferença entre o valor de um ponto e a média, pelo desvio padrão. Compara-se o resultado, em valor absoluto, com o limite da tabela, (interpolado, se necessário). Se for maior, o ponto é discrepante:

$z = |(x - \bar{x}) / \sigma|$

x é discrepante se $z > z_d$

Os pontos da amostra devem ser ordenados. Examina-se inicialmente o ponto mais afastado da média. Se ele for discrepante, é eliminado, observando-se o seguinte.

O ponto discrepante não deve ser eliminado, se a diferença em relação à média for menor que a diferença admissível *d*. A diferença admissível *d* é a maior diferença que ainda não é considerada significativa (ver o item 3.7 da terceira parte).

Eliminado o ponto, recalcula-se a média e o desvio padrão. O processo deve ser repetido, até que todos os pontos discrepantes sejam eliminados, ou até que tenham sido eliminados 1/3 dos pontos. Devem restar, na amostra, pelo menos 2/3 dos pontos originais. A média e o desvio padrão, calculados com a amostra remanescente, representam melhor os valores verdadeiros.

2.6 – COMPARAÇÃO ENTRE O MÉTODO DO TESTE t E O MÉTODO PROPOSTO

a) avaliação do método do teste t de Student

Neste método, como visto, $x_d = \bar{x} \pm t\,\sigma\,\sqrt{(1/n + 1)}$.

Para avaliar este método, foram imaginadas algumas amostras aproximadamente normais, para as quais foram calculados os valores de x_d, conforme o anexo 3.

Os valores de x_d foram então avaliados apenas com base no bom senso.

Considerando uma amostra de 3 pontos (n = 2):

$x_1 = 8$, $x_2 = 10$, $x_3 = 12$

usando x_1 e x_2 obtemos: $\bar{x} = 9$ $\sigma = 1{,}414$ $\sqrt{(1/n + 1)} = 1{,}225$

com $\alpha=0{,}05$: $x_d = 9 + 6{,}31 \times 1{,}414 \times 1{,}225 = 19{,}9$ (alto, quase o dobro da média)

com $\alpha=0{,}025$: $x_d = 9 + 12{,}71 \times 1{,}414 \times 1{,}225 = 31{,}0$ (alto demais)

Considerando uma amostra de 4 pontos (n = 3):

$x_1 = 8$, $x_2 = 10$, $x_3 = 12$, $x_4 = 14$

com os três primeiros, obtemos: $\bar{x} = 10$ $\sigma = 2$ $\sqrt{(1/n + 1)} = 1{,}155$

com $\alpha=0{,}05$: $x_d = 10 + 2{,}92 \times 2 \times 1{,}155 = 16{,}7$ (razoável)

com $\alpha=0{,}025$: $x_d = 10 + 4{,}30 \times 2 \times 1{,}155 = 19{,}9$ (razoável)

Considerando uma amostra de 8 pontos (n = 7):

x:	7	9	11	13
f:	1	3	3	1

excluindo o último ponto, temos: $\bar{x} = 9{,}57$ $\sigma = 1{,}512$ $\sqrt{(1/n + 1)} = 1{,}069$

com $\alpha=0{,}05$: $x_d = 9{,}57 + 1{,}94 \times 1{,}512 \times 1{,}069 = 12{,}7$ (baixo, menor que o último ponto)

com $\alpha=0{,}025$: $x_d = 9{,}57 + 2{,}45 \times 1{,}512 \times 1{,}069 = 13{,}5$ (razoável)

Considerando uma amostra de 9 pontos (n = 8):
x: 6 8 10 12 14
f: 1 2 3 2 1
excluindo o último ponto, temos: $\bar{x}$ = 9,5 σ = 2,07 √(1/n + 1) = 1,061
com α=0,05: x_d = 9,5 + 1,89 × 2,07 × 1,061 = 13,6 (baixo, menor que o último ponto)
com α=0,025: x_d = 9,5 + 2,36 × 2,07 × 1,061 = 14,7 (razoável)

Constata-se que o método do teste t, embora teoricamente fundamentado, nem sempre apresenta resultados condizentes com o bom senso.
Nas amostras menores, os valores de x_d são muito altos. Como vimos anteriormente, o limite discrepante inferior não deve ser menor que zero. Logo, sendo a distribuição simétrica, o limite discrepante superior não poderá ser maior que o dobro do valor médio da variável. No primeiro exemplo analisado acima, a média dos 3 pontos é 10.
O valor máximo não deveria ultrapassar 20. Portanto o limite x_d = 31, calculado pelo teste t, é alto demais (um problema semelhante ocorre no método de Grubbs, analisado no anexo 5). Nas amostras maiores o método do teste t acusa, como discrepantes, pontos pertencentes à amostra normal, o que é um contra-senso.

b) avaliação do método proposto
O método proposto foi avaliado considerando os mesmos exemplos acima. Os valores de x_d foram determinados usando a tabela 3. O inverso da tabela requer uso de tentativas; aumenta-se o valor do último ponto de cada amostra até que seja atingido o limite z_d. Foram obtidos os resultados abaixo.

amostra de			
	3 pontos:	para z_d = 1,121, x_d = 16,0	(razoável)
	4 pontos:	para z_d = 1,391, x_d = 18,1	(razoável)
	8 pontos:	para z_d = 1,822, x_d = 13,9	(razoável)
	9 pontos:	para z_d = 1,881, x_d = 15,3	(razoável)

Vê-se que o método proposto não apresenta os problemas constatados no método do teste t. Nas amostras menores, os limites são inferiores ao dobro do valor médio da variável. Nas amostras maiores, os limites estão acima dos valores máximos das amostras.

c) justificativas do método proposto
Comparado com o método do teste t, o método proposto apresenta as seguintes vantagens:
–A hipótese adotada (de médias diferentes) é mais adequada para avaliar pontos discrepantes.
–Com grandezas por natureza positivas, o método não aceita valores negativos ou muito altos (mais que o dobro) em relação à média.
–O método considera o aumento da amplitude da amostra com o tamanho da mesma.
–É muito fácil de usar.
O método proposto também foi comparado com outros métodos. As vantagens do método proposto em relação aos métodos de Chauvenet, de Grubbs e ao "box&whisker" ficam evidentes nos anexos 4, 5 e 6.

2.7 – EXEMPLO

Pretende-se adquirir um eletrodoméstico com preço aproximado de R$ 1000,00. Neste nível de custo, considera-se que uma diferença de R$ 20,00 não é significativa (*d* = 20). Foram obtidos 3 preços:

a) R$ 800,00, R$ 1000,00 e R$ 1700,00. O preço mais alto é discrepante?
Aplicando o método proposto, temos:
média = 1166,67
desvio padrão = 472,58
diferença = 1700,00 - 1166,67 = 533,33
z = 533,33 / 472,58 = 1,128
da tabela 3, para N = 3, z_d = 1,121
como z é maior que z_d, o ponto é discrepante. A diferença é maior que *d*, portanto significativa, confirmando que o ponto deve ser eliminado.

Observação: Pelo método proposto, o valor discrepante superior é R$ 1600,00.
No método de Grubbs, o valor discrepante superior seria R$ 2810,00. Pelo teste t de Student (nível α=0,025) seria ainda mais alto, R$ 3100,00. Portanto, em ambos, o preço mais alto não seria considerado discrepante.

Observação: O valor de z deve ser calculado, com precisão, até a terceira casa decimal.

b) R$ 975,00, R$ 1000,00, R$ 1000,00. O preço mais baixo é discrepante?
Aplicando o método proposto, temos:
média = 991,67
desvio padrão = 14,43
diferença = 991,67 - 975,00 = 16,67
z = 16,67 / 14,43 = 1,155
da tabela 3, para N = 3, z_d = 1,121
como z é maior que z_d, o ponto é discrepante. Porém a diferença é menor que *d*; o ponto não deve ser eliminado.

Observação: Quando uma amostra tem vários valores iguais, qualquer ponto um pouco diferente tenderá a ser indicado como discrepante; neste caso, o critério decisivo é a diferença admissível *d*.

ANEXOS RELATIVOS À SEGUNDA PARTE

ANEXO 3 – Exemplos de cálculo de x_d pelo método do teste t
ANEXO 4 – O critério de Chauvenet
ANEXO 5 – Comparação com o método de Grubbs
ANEXO 6 – O método "box&whisker"

3 – TERCEIRA PARTE: TAMANHO DA AMOSTRA

Por Karl Grube e Gerhard Grube, Junho de 2012.

3.1 – INTRODUÇÃO E RESUMO

Em qualquer experimento, enfrenta-se o problema de determinar o tamanho da amostra ("sample size"). Quanto maior a amostra, mais preciso será o resultado do experimento. Por outro lado, o custo de obtenção da amostra aumenta.
O método indicado na literatura para determinar o tamanho da amostra, baseado na diferença entre a média da amostra e a média verdadeira, pode resultar em amostras muito grandes. Se o custo de obtenção da amostra é elevado, a amostragem pode se revelar inviável.
Nesta terceira parte propomos um método alternativo que indica tamanhos mais razoáveis. No método aqui proposto, limita-se o tamanho da amostra quando um ponto hipotético adicional, arbitrariamente alto, não pode alterar significativamente a média e esta fica praticamente estável. Com este critério, elaborou-se uma tabela, apresentada no Item 3.4, que permite determinar o tamanho da amostra, em função da relação entre o desvio padrão e a diferença admissível.
Sem deixar de ser confiável, o método proposto resulta em tamanhos bem menores que o da literatura. Relativamente poucos pontos são suficientes para estabilizar a média. Deste modo, sendo altos os custos, a amostragem pode ser viabilizada.
O método foi verificado por simulação com amostras aleatórias extraídas de uma distribuição normal (anexo 7).
Um aspecto não considerado no método da literatura é que, aumentando o tamanho da amostra, a contribuição de cada ponto adicional para a precisão do resultado diminui, até atingir o ponto em que se torna desprezível ou nula, ou seja, que existe um tamanho máximo para a amostra. Esta questão é examinada no anexo 8.

3.2 – MÉTODO DA LITERATURA

Conforme a literatura, o tamanho N da amostra pode ser estabelecido se forem conhecidos, ao menos aproximadamente, o desvio padrão σ da população e o erro e (diferença entre a média da amostra e a média verdadeira). O tamanho da amostra aumenta com a relação σ/e.
A literatura apresenta a seguinte equação:

$$N = (z\,\sigma/e)^{\wedge}2 \qquad \text{equação [8]}$$

em que z = afastamento da média conforme o nível de confiança desejado.
Num exemplo didático apresentado na Ref. 2 (pág. 201), para um nível de confiança de 90% (z = 1,65), e = 1 e σ = 10, o tamanho requerido é N = 273.
Se os custos de amostragem forem elevados, um número tão alto provavelmente representaria um problema intransponível na realização de um projeto.

3.3 – DESENVOLVIMENTO DO MÉTODO PROPOSTO

Aumentando-se o tamanho de uma amostra, cada ponto adicional tem uma influência menor sobre o valor da média calculada. Por exemplo, o peso de um terceiro ponto no cálculo da média é 1/3, já de um décimo ponto acrescentado à amostra vai ser apenas 1/10, e assim por diante. As flutuações da média calculada tendem a diminuir e a média tende a ficar estável. Atingida certa estabilidade, há pouco benefício em prosseguir com a amostragem. Este pode ser um critério válido para determinar o tamanho da amostra.
Com relação a este tipo de abordagem, Pillar (Ref. 4, pág. 6) aponta o problema de que a percepção de estabilidade é afetada pela seqüência real dos pontos (uma flutuação grande no início dá a impressão de estabilidade; já no final, a impressão é inversa).
No método proposto este problema é evitado, porque o tamanho da amostra é definido somente pela diferença causada pelo ponto, não importando a sua posição; como se verá, o valor de N deduzido abaixo independe da ordem na qual o ponto extremo x_d é agregado à amostra.

a) Equação do tamanho da amostra
A idéia básica do método é definir o tamanho da amostra quando um ponto hipotético adicional, arbitrariamente alto, não pode mais alterar significativamente a média. Com isto, garante-se que a média fica relativamente estável. O tamanho N da amostra necessário para que isto ocorra é deduzido a seguir.
Sejam $\overline{x}$ a média e σ o desvio padrão calculados para uma amostra de tamanho N. Seja x_d um ponto extremo escolhido arbitrariamente, igual a
$x_d = \overline{x} + z'\sigma$
sendo z' definido pelo nível de confiança desejado. A nova média, incluindo o ponto adicional x_d, é
$\overline{x}' = (N \times \overline{x} + \overline{x} + z'\sigma) / (N+1)$
A alteração na média, causada pelo ponto adicional, não deverá ser maior que *d,* a diferença admissível em relação à média calculada:
$\overline{x}' - \overline{x} \leq d$
Substituindo e transformando, obtemos:

$N \geq z'\sigma/d - 1$ equação [9]

Verifica-se que o valor de N independe da ordem em que o ponto x_d foi agregado à amostra. Nesta expressão, o termo $z'\sigma/d$ não é elevado ao quadrado e resultará sempre em valores de N menores que a equação [8].
Para o exemplo citado no item 3.2 acima, o tamanho da amostra para $d = 1$, $\sigma = 10$, considerando $z' = 2{,}58$ (correspondente a um nível de confiança de 99%), passa a ser
$N = 2{,}58 \times 10 - 1 = 24{,}8 \approx 25$
que é um tamanho de amostra bem mais razoável.

b) Intervalo de confiança da média
À medida que o tamanho da amostra aumenta, a diferença *d* é atingida muito antes do erro *e*. Assim, com relativamente poucos pontos, já se obtém uma média estável, com a qual se pode estimar a média verdadeira.

Para estimar a média verdadeira, define-se um erro e' com o novo tamanho da amostra, empregando a equação [8] transformada:
$e' = z \sigma / \sqrt{N}$.
o intervalo de confiança da média verdadeira é dado por

$\mu = \bar{x} \pm e' = \bar{x} \pm z \sigma / \sqrt{N}$ equação [10]

No exemplo em questão, para um nível de confiança de 90% ($z = 1,65$) e N=25:
$e' = 1,65 \times 10 / \sqrt{25} = 3,3$
A nova estimativa da média verdadeira é, portanto

$\mu = \bar{x} \pm 3,3$

Assim, aplicando o método proposto, reduziu-se o tamanho da amostra de 273 para 25 pontos; por outro lado, o intervalo de confiança (90%) da média verdadeira é 3,3 vezes maior. O intervalo de confiança maior é o preço que se paga pela economia nos custos de amostragem. Entretanto, o importante é que se pode ter uma elevada confiança (99%) de que o valor calculado da média não mudará por uma diferença maior que *d*.

c) Efeito da amplitude da amostra
O valor de N é dado pela equação [9]. Explicitada para o valor de d/σ, a equação fica

$d/\sigma = z' / (N+1)$ equação [11]

Rigorosamente, o valor a ser escolhido para z' na equação [11] não é independente do tamanho da amostra. Tendo em vista que a amplitude de uma amostra aumenta com o tamanho da mesma, o valor de z' também deve aumentar. Verificou-se, pela simulação realizada (anexo 7), que um fator de correção é necessário para considerar este efeito. O fator de correção, determinado no anexo 7 (item a) é:

$h = 1 + 0,001\ N$

Escolhendo um nível de confiança de 99% ($z' = 2,58$) e agregando o fator de correção, a equação final para o valor de d/σ é:

$d/\sigma = 2,58/(N+1) \times (1 + 0,001\ N)$ equação [12]

3.4 – TAMANHO DA AMOSTRA – MÉTODO PROPOSTO

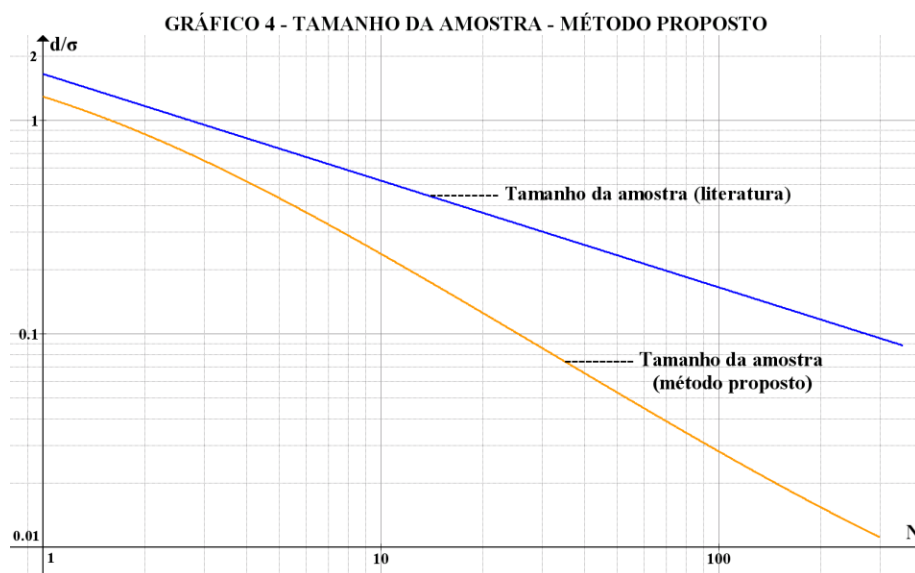
A tabela 4 abaixo apresenta os valores de d/σ em função do tamanho da amostra, determinados conforme a equação [12]. Com esta tabela, dado um valor de d/σ, pode-se determinar o tamanho N da amostra.

TABELA 4 – TAMANHO DA AMOSTRA – MÉTODO PROPOSTO

N	*d*/σ	N	*d*/σ	N	*d*/σ	N	*d*/σ
1	≥1,29	12	0,201	35	0,074	90	0,031
2	0,862	14	0,174	40	0,065	100	0,028
3	0,647	16	0,154	45	0,059	110	0,026
4	0,518	18	0,138	50	0,053	120	0,024
5	0,432	20	0,125	55	0,049	130	0,022
6	0,371	22	0,115	60	0,045	140	0,021
7	0,325	24	0,106	65	0,042	150	0,020
8	0,289	26	0,098	70	0,039	160	0,019
9	0,260	28	0,091	75	0,036	180	0,017
10	0,237	30	0,086	80	0,034	200	0,015

N: número de pontos da amostra
d/σ: diferença admissível / desvio padrão estimado

Numa planilha (Open Office, Excel 2010), o valor de *d*/σ pode ser obtido mediante a expressão: =(2,58/(D13+1))*(1+ 0,001*D13)
em que D13 é a célula que contém N (número de pontos da amostra).

Os valores da equação [12] estão representados pela linha amarela no gráfico 4 a seguir. Para comparação, a linha reta em azul indica os valores pelo método da literatura (para um nível de confiança de 90%). Fica evidente a grande redução nos tamanhos de amostra, proporcionada pelo método proposto.

N: tamanho da amostra
d/σ: diferença admissível / desvio padrão estimado
Obs: gráfico log-log

3.5 – APLICAÇÃO DO MÉTODO PROPOSTO

Para aplicar o método, deve-se ter uma estimativa da diferença admissível e uma estimativa do desvio padrão. Os tamanhos indicados na tabela 4 acima são os mínimos recomendados para tornar estável a média calculada. Nada impede que sejam usadas amostras maiores, por exemplo, quando a população amostrada é heterogênea, visando garantir que a amostra seja representativa.
Após realizar a amostragem com os N pontos da tabela 4, deve ser verificada a existência de pontos discrepantes, conforme a segunda parte deste trabalho. Os pontos discrepantes devem ser eliminados e substituídos por outros, completando o tamanho N requerido.
A média calculada é a melhor estimativa da média verdadeira, cujo intervalo de confiança pode ser determinado conforme a equação [10] do item 3.3.

Cabe aqui mencionar outro problema apontado por Pillar (Ref. 4, pág. 6), de que a precisão desejada pode ser atingida antes da média se tornar estável (o que levaria a interromper a amostragem cedo demais). A observação não se aplica ao método proposto, porque o tamanho N é definido pela diferença máxima possível, não pela real, que é menor. Assim, mesmo que a relação d/σ desejada já tenha sido atingida, deve-se prosseguir a amostragem até chegar ao valor N recomendado na tabela 4, quando a média fica estável.

3.6 – VERIFICAÇÃO DO MÉTODO PROPOSTO

Para verificar o método, foram realizadas simulações de amostragens aleatórias de uma distribuição normal com média $\mu = 10$ e desvio padrão $\sigma = 1$. Foram obtidas quatro mil amostras para diversos valores de N. Os resultados são mostrados no anexo 7. Todos os valores de d/σ obtidos nas simulações encontram-se abaixo dos indicados na tabela 4 do item 3.4. Isto confirma que os tamanhos de amostra recomendados pelo método proposto são adequados para os casos reais. O nível de confiança do método é superior a 99%. Os valores medianos de d/σ obtidos nas simulações variam aproximadamente proporcionais a 1/N.

3.7 – ESTIMATIVA DA RELAÇÃO d/σ

O tamanho da amostra deve ser determinado em função da relação entre a diferença admissível d e o desvio padrão estimado da população, σ.
Em geral é possível obter uma estimativa razoável do desvio padrão, analisando o comportamento esperado da variável. Tendo-se uma idéia da faixa de variação, pode-se admitir, grosso modo, que ela equivale a 5 ou 6 desvios padrão. Após realizar alguns testes (pelo menos três), o valor do desvio padrão poderá ser recalculado, revisando-se, se necessário, o tamanho da amostra.
A diferença admissível d é a maior diferença, na média calculada, que ainda não é considerada significativa, face os objetivos da pesquisa. A sua estimativa pode ser bem difícil; é preciso avaliar o efeito da diferença sobre os objetivos finais da pesquisa, que

muitas vezes não são bem conhecidos. Por exemplo, no desenvolvimento de um processo industrial, seria necessário estimar o efeito da diferença admissível sobre o resultado econômico do processo.
Muitas vezes a única saída é considerar a diferença admissível igual a um erro aceitável no valor da variável. Não se deve escolher um erro pequeno demais. Erros menores que 1% são difíceis de serem obtidos. Em pesquisa tecnológica, às vezes se aceita erros de 10% ou mais. De qualquer modo, uma vez definida a diferença admissível, o sistema de medição deve ser escolhido com a precisão adequada. O erro de medição deve ser bem menor que a diferença admissível.

3.8 – TAMANHO MÁXIMO DA AMOSTRA

As relações d/σ diminuem com o tamanho da amostra, até se tornarem desprezíveis ou nulas. Neste ponto, foi atingido um tamanho máximo razoável da amostra; não compensa acrescentar mais pontos. No anexo 8 apresentamos razões para afirmar que, quando os custos de amostragem são significativos, dificilmente se justificam amostras com mais de 30 pontos; outra conclusão é que não há interesse em amostras com mais de 200 pontos, mesmo que os custos de amostragem sejam muito baixos.

3.9 – EXEMPLO

Deseja-se pesquisar o preço de um equipamento industrial cujo valor, numa primeira estimativa, é de R$ 150.000,00, com uma faixa de variação entre R$ 100.000,00 e R$ 200.000,00.
Quantas propostas deverão ser solicitadas? Em geral, neste tipo de pesquisa, os custos não são desprezíveis; a elaboração e a análise das propostas sempre exigem um tempo considerável. Portanto, deve-se procurar a quantidade mínima necessária.
O desvio padrão aproximado é (200.000 - 100.000) / 5 = 20.000
A diferença admissível *d* deve ser definida conforme o objetivo da pesquisa. Se, por exemplo, o objetivo for uma estimativa preliminar de custos, uma diferença de 10% do preço esperado é aceitável; portanto,
$d = 150.000 \times 0,10 = 15.000$
$d/\sigma = 15.000 / 20.000 = 0,75$
Da Tabela 4, obtemos N = 3
Assim, para uma estimativa preliminar de custos, bastam três propostas.

Observação: Embora esta quantidade não seja grande, pode-se ter confiança que uma proposta adicional não irá alterar o valor médio calculado por uma diferença maior que ±15.000. Se o valor médio calculado é 150.000, uma diferença maior só poderia ser causada por uma proposta adicional maior que 210.000 ou menor que 90.000, portanto fora da faixa estimada da variável.

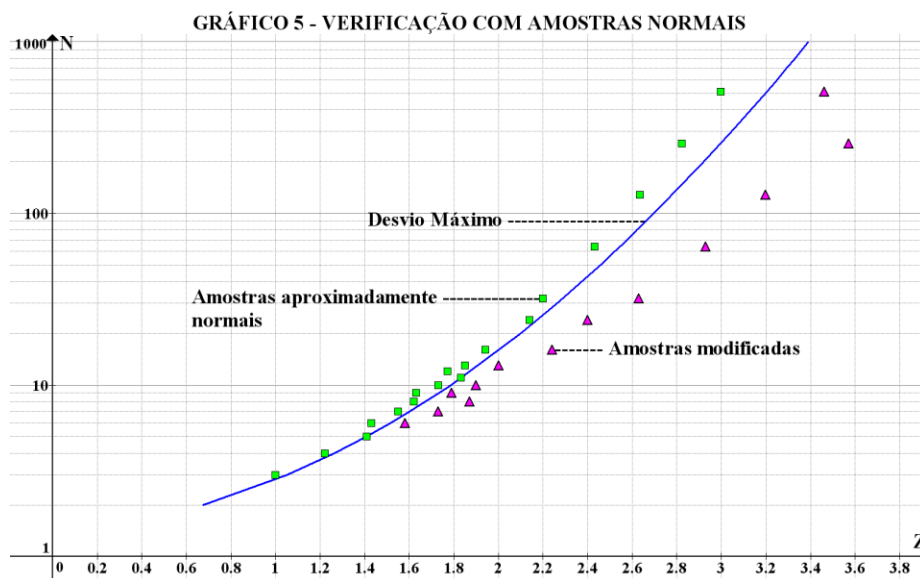
ANEXOS RELATIVOS À TERCEIRA PARTE

Anexo 7 – Simulação de amostragens reais
Anexo 8 – Limites máximos para o tamanho da amostra

ANEXO 1 (da primeira parte)

VERIFICAÇÃO COM ALGUMAS AMOSTRAS NORMAIS

A equação para o desvio máximo foi verificada utilizando-se algumas amostras hipotéticas aproximadamente normais. Para estas amostras foram calculados o desvio padrão e o desvio máximo em relação à média (tabela 5). Estes pontos foram locados no gráfico 5 (quadrados verdes). Como esperado, os pontos situam-se em geral à esquerda da curva teórica.

Na tabela 6 algumas destas amostras foram modificadas, deslocando alguns pontos para o centro, sem alterar a simetria. Com isto, as amostras passam a apresentar um pico mais alto que o normal e os pontos extremos tendem a ficar discrepantes. Locados no gráfico (triângulos roxos), verifica-se que se situam à direita da curva de desvios máximos, conforme esperado.

N: número de pontos da amostra

Z: afastamento = $(x - \overline{x}) / \sigma$

TABELA 5 – AMOSTRAS APROXIMADAMENTE NORMAIS (média = 10,0)

N	x / freq										σ	z
3	8/1	10/1	12/1								2,000	1,00
4	8/1	10/2	12/1								1,633	1,22
5	8/1	10/3	12/1								1,414	1,41
6	7/1	9/2	11/2	13/1							2,098	1,43
7	6/1	8/1	10/3	12/1	14/1						2,582	1,55
8	7/1	9/3	11/3	13/1							1,852	1,62
9	8/1	9/2	10/3	11/2	12/1						1,225	1,63
10	6/1	8/2	10/4	12/2	14/1						2,309	1,73
11	6/1	8/2	10/5	12/2	14/1						2,191	1,83
12	6/1	8/3	10/4	12/3	14/1						2,256	1,77
13	6/1	8/3	10/5	12/3	14/1						2,160	1,85
16	6/1	8/4	10/6	12/4	14/1						2,066	1,94
24	6/1	8/6	10/10	12/6	14/1						1,865	2,14
32	5/1	7/5	9/10	11/10	13/5	15/1					2,272	2,20
64	7/1	8/6	9/15	10/20	11/15	12/6	13/1				1,234	2,43
128	3/1	5/7	7/21	9/35	11/35	13/21	15/7	17/1			2,656	2,64
256	6/1	7/8	8/28	9/56	10/70	11/56	12/28	13/8	14/1		1,417	2,82
512	1/1	3/9	5/36	7/84	9/126	11/126	13/84	15/36	17/9	19/1	3,003	3,00

N: número de pontos da amostra
x / freq: valor / freqüência
σ: desvio padrão da amostra
z: afastamento do maior ponto (desvio máximo)

TABELA 6 – AMOSTRAS MODIFICADAS (média = 10,0)

N	x / freq										σ	z
6	8/1	10/4	12/1								1,265	1,58
7	8/1	10/5	12/1								1,155	1,73
8	8/1	10/6	12/1								1,069	1,87
9	8/1	9/1	10/5	11/1	12/1						1,118	1,79
10	6/1	8/1	10/6	12/1	14/1						2,108	1,90
13	6/1	8/2	10/7	12/2	14/1						2,000	2,00
16	6/1	8/2	10/10	12/2	14/1						1,789	2,24
24	6/1	8/4	10/14	12/4	14/1						1,668	2,40
32	5/1	7/2	9/13	11/13	13/2	15/1					1,901	2,63
64	7/1	8/3	9/12	10/32	11/12	12/3	13/1				1,024	2,93
128	3/1	5/3	7/15	9/45	11/45	13/15	15/3	17/1			2,188	3,20
256	6/1	7/4	8/12	9/60	10/102	11/60	12/12	13/4	14/1		1,120	3,57
512	1/1	3/5	5/20	7/84	9/146	11/146	13/84	15/20	17/5	19/1	2,601	3,46

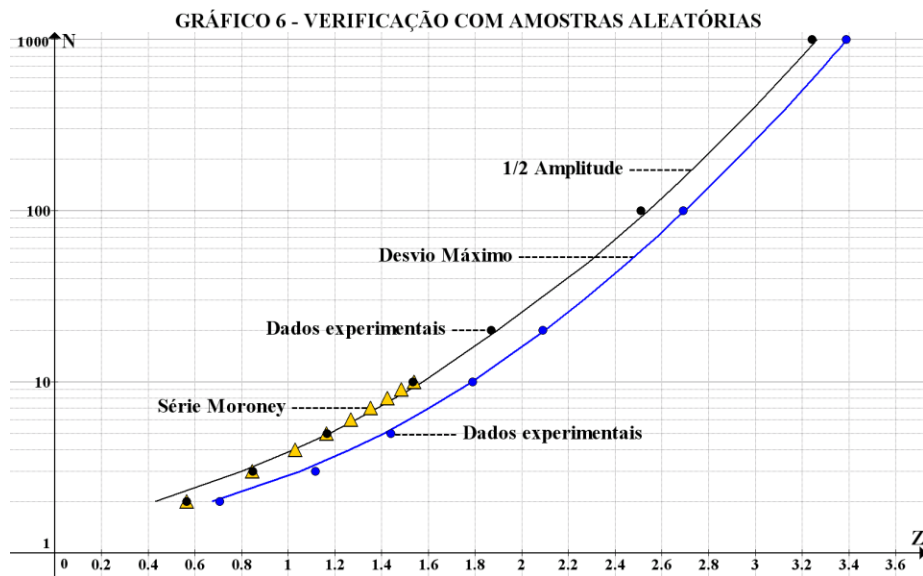
N: número de pontos da amostra
x / freq: valor/ freqüência
σ: desvio padrão da amostra
z: afastamento do maior ponto (desvio máximo)

ANEXO 2 (da primeira parte)

VERIFICAÇÃO COM AMOSTRAS ALEATÓRIAS

Para uma verificação mais exaustiva obteve-se amostras aleatórias da distribuição normal (média 10, desvio padrão 1), usando a função =INV.NORM(ALEATÓRIO();10;1) da planilha Excel 2010. Foram obtidas dez mil amostras de cada tamanho para os seguintes tamanhos: 2, 3, 5, 10, 20, 100 e 1000.

a) Os desvios máximos esperados foram estimados da seguinte maneira:
Para um determinado valor de N, determinou-se:
–a média de cada uma das amostras
–o desvio padrão da amostra
–a diferença entre o maior ponto da amostra e a média da amostra
–a diferença entre a média da amostra e o menor ponto
–selecionou-se a maior das duas diferenças acima
–dividiu-se a diferença selecionada pelo desvio padrão da amostra
A mediana dos dez mil resultados para cada tamanho foi locada no gráfico 6 contra o valor de N (bolinhas azuis). Neste mesmo gráfico, a linha azul representa os valores teóricos. Os valores experimentais se situam próximos da linha teórica.

b) A amplitude também foi verificada, através da mesma série, calculando a metade da diferença entre o maior ponto da amostra e o menor. A média dos dez mil resultados também foi locada no gráfico 6.
Os pontos desta série (bolinhas pretas) se localizam em geral próximo à linha teórica (em preto); a maior diferença ocorre nas amostras de 2 pontos; nestas, os valores experimentais estão cerca de 0,1σ acima do valor teórico.
Moroney (Ref. 5, pág. 155), publicou valores da amplitude para N de 2 até 10, sem indicar como foram obtidos. Estes estão representados no gráfico pelos triângulos amarelos. A proximidade com as bolinhas pretas confirma o procedimento experimental aqui adotado.

N: número de pontos da amostra

Z: afastamento= $(x - \bar{x}) / \sigma$

ANEXO 3 (da segunda parte)

EXEMPLOS DE CÁLCULOS PELO MÉTODO DO TESTE t

Cálculo dos valores de x_d pelo método do teste t de Student

n	g.l.	√(1/n + 1)	t(α=0,025)	xd/σ	t(α=0,05)	xd/σ
2	1	1,2247	12,706	15,561	6,314	7,733
3	2	1,1547	4,303	4,969	2,920	3,372
4	3	1,1180	3,182	3,557	2,353	2,631
7	6	1,0690	2,447	2,616	1,943	2,077
8	7	1,0607	2,365	2,509	1,895	2,010
19	18	1,0260	2,101	2,156	1,734	1,779
41	40	1,0121	2,021	2,045	1,684	1,704
100	99	1,0050	1,987	1,997	1,663	1,671

n: número de pontos da amostra, menos um

g.l.: graus de liberdade

t(α): valor do t de Student para o nível de significância α

$x_d/\sigma = t\sqrt{1/n + 1}$

ANEXO 4 (da segunda parte)

O CRITÉRIO DE CHAUVENET

No Apêndice D do livro de Vuolo (Ref. 3), é apresentado o critério de Chauvenet, na forma de uma tabela de limites discrepantes em função do tamanho da amostra. Eles estão muito próximos dos limites determinados (para amostras maiores) nesta segunda parte. Chauvenet definiu como limites, intervalos simétricos da distribuição normal com a probabilidade

$p = 1 - 1/2\ N$

Abaixo estão relacionados alguns dos limites constantes na Ref. 3 (a tabela só começa com N = 8), em comparação com os aqui obtidos:

N	Chauvenet	Critério 2
8	1,86	1,82
10	1,96	1,93
12	2,04	2,01
15	2,13	2,11
. . . .	. . . .	. . . .
200	3,02	3,02
500	3,29	3,29
1000	3,48	3,48

N: número de pontos da amostra

A coincidência é notável, considerando que no presente trabalho a probabilidade para o limite discrepante foi obtida por um raciocínio diferente, resultando também numa fórmula bastante diferente:

$p = 0{,}6089\^(1/n)$ onde $n = N - 1$

Para menos de 8 pontos, os limites de Chauvenet são amplos demais (talvez por esta razão, Vuolo os omitiu), apresentando problemas semelhantes aos encontrados no teste t de Student e no método de Grubbs, analisado no anexo 5.
O método proposto, indicando limites mais adequados para os tamanhos menores, pode ser considerado uma complementação útil do critério de Chauvenet.

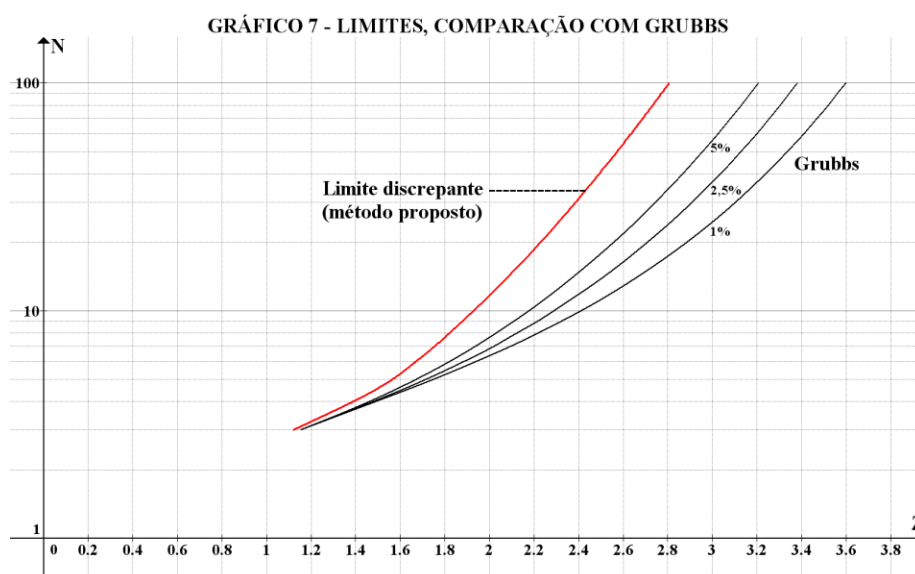
ANEXO 5 (da segunda parte)

COMPARAÇÃO COM O MÉTODO DE GRUBBS

a) comparação entre os limites
Um método bastante divulgado na literatura foi desenvolvido por Grubbs. Esse autor apresenta, na Ref. 6, os procedimentos que devem ser seguidos para a identificação de pontos discrepantes. Analisaremos aqui os limites recomendados na pág. 4. Estes limites são apresentados na forma de tabela em função de N, dos quais reproduzimos, abaixo, apenas os pontos iniciais. Para comparação, colocamos os limites desenvolvidos no presente trabalho. Os dados mais completos estão representados no gráfico 7, a seguir.

TABELA 7 – LIMITES DISCREPANTES (COMPARAÇÃO COM GRUBBS)

N	Limites de Grubbs			Método
	1%	2,5%	5%	Proposto
3	1,15	1,15	1,15	1,121
4	1,49	1,48	1,46	1,391
5	1,75	1,71	1,67	1,565
6	1,94	1,89	1,82	1,672
7	2,10	2,02	1,94	1,754

N: número de pontos da amostra

N: número de pontos da amostra

Z: afastamento, $(x - \bar{x}) / \sigma$

%: nível de significância

Observa-se de imediato, que todos os limites de Grubbs são bem mais altos que os encontrados aqui (a proximidade para as amostras menores é só aparente; como se verá adiante, uma pequena diferença entre os limites implica, na realidade, em uma grande diferença nos valores discrepantes). Conseqüentemente, os limites de Grubbs tendem a acusar menos pontos discrepantes. Uma simulação mostrou que o critério de Grubbs (nível 5%) rejeita menos de 1/3 dos pontos rejeitados pelo método proposto.

Tendo em vista que as curvas são semelhantes, a razão poderia estar nos níveis de confiança adotados. No método proposto, os limites são bastante restritivos, porque um foco principal do método é o tratamento adequado de grandezas que, por natureza, só podem ser positivas. Os limites restritivos diminuem a probabilidade de aceitar valores negativos, principalmente nas amostras menores.

b) comparação entre os valores discrepantes
A diferença entre os dois métodos fica mais evidente quando se calcula os valores discrepantes x_d decorrentes dos respectivos limites. Para isto, recorremos a algumas das amostras aproximadamente normais utilizadas na primeira parte deste trabalho (anexo 1, tabela 5). Todas as amostras apresentam um valor médio igual a 10, com desvios padrão variando entre 1,41 e 2,58. Com os limites da tabela acima, foram determinados (por tentativas) os valores discrepantes x_d. Os resultados constam na tabela a seguir.

TABELA 8 – VALORES DISCREPANTES (COMPARAÇÃO COM GRUBBS)

N	Valores Discrepantes Inferiores				Valores Discrepantes Superiores			
		Grubbs		Método		Grubbs		Método
	1%	2,5%	5%	Proposto	1%	2,5%	5%	Proposto
3	-8,1	-8,1	-8,1	4,0	28,1	28,1	28,1	16,0
4	-5,6	-0,8	2,7	6,0	25,6	20,8	17,3	14,0
5	1,4	4,2	5,5	7,0	18,6	15,8	14,5	13,0
6	-0,6	1,6	3,4	5,4	20,6	18,4	16,6	14,6
7	-1,6	0,9	2,4	4,6	21,6	19,1	17,6	15,4

N: número de pontos da amostra

Conforme observado acima, vê-se que, para N = 3, a diferença entre o limite de Grubbs e o do método proposto (1,15 contra 1,121, tabela 7), aparentemente pequena, resulta numa diferença muito grande nos valores discrepantes superiores (28,1 contra 16,0, tabela 8).
Se a grandeza em questão é positiva, são inadmissíveis valores negativos e, por simetria, os valores maiores que 20. Constata-se que os limites de Grubbs resultam, em alguns casos, em valores discrepantes negativos e outros maiores que 20.
É preciso ressaltar que, numa distribuição normal, valores negativos só podem ocorrer, com probabilidade significativa, se o coeficiente de dispersão (σ/μ) for maior que 1/3. Como nas amostras analisadas os coeficientes de dispersão estão entre 0,141 e 0,258, os valores discrepantes negativos não se justificam. Os limites de Grubbs são amplos demais.
Pelo método proposto, nas amostras analisadas, nenhum valor discrepante é negativo ou maior que 20 (tabela 8). A maioria das variáveis com que se lida em engenharia são grandezas positivas; valores negativos são impossíveis. Somos da opinião que um método para identificação de pontos discrepantes só terá aplicação geral se considerar adequadamente este fato.

c) utilização da transformação log-normal
Poderia ser contraposto, ao que foi dito acima, que os limites negativos sempre poderão ser evitados transformando-se a distribuição amostral em uma distribuição

log-normal. Nesta transformação (que, rigorosamente, só deveria ser usada para tornar normais algumas distribuições assimétricas), os valores da variável são substituídos pelos seus logaritmos. Demonstraremos a seguir que, quando os limites são muito amplos, este recurso não é satisfatório.
Tomamos como exemplo a amostra de 3 pontos do anexo 1, tabela 5:
$x_1=8$, $x_2=10$, $x_3=12$
Analisemos, inicialmente, os efeitos da transformação sobre os limites de Grubbs.
Para esta amostra, os valores discrepantes (sem transformação) conforme Grubbs constam da tabela 8 acima. O valor superior é 28,1 e o inferior é negativo, -8,1.
A transformação é feita substituindo os valores de x da amostra pelos seus logaritmos; obtemos:
$x_1=2{,}0794$, $x_2=2{,}3026$, $x_3=2{,}4849$
para achar o valor discrepante superior, são necessárias tentativas. Aumenta-se o valor de x_3 até atingir o limite discrepante especificado (1,15, segundo Grubbs).
obtém-se $x_d=4{,}33$
de modo semelhante, o valor discrepante inferior é encontrado diminuindo-se o valor de x_1 até atingir o limite.
obtém-se $x_d=0{,}65$
operando a transformação inversa (antilog x_d), obtém-se os valores discrepantes:

superior: 75,9
inferior: 1,9

Vemos que a transformação log-normal eliminou o valor negativo, mas os resultados não são razoáveis. O valor discrepante inferior deixou de ser negativo, mas ainda é muito baixo. O valor discrepante superior ficou desproporcionalmente alto.
A transformação log-normal não resolve os problemas causados pelos limites muito amplos de Grubbs.
Com o método proposto, não ocorrem estes problemas. Sem transformação, os valores discrepantes já são positivos (superior: 16,0, inferior: 4,0, tabela 8).
Transformando a amostra, o limite do método proposto (1,121) resulta nos valores discrepantes transformados:
superior $x_d=2{,}97$
inferior $x_d=1{,}75$
a transformação inversa dá os valores discrepantes:

superior: 19,5
inferior: 5,8

Vê-se que, mesmo aplicando a transformação, os valores discrepantes pelo método proposto continuam perfeitamente razoáveis.

ANEXO 6 (da segunda parte)

O MÉTODO "BOX&WHISKER"

Este método é mencionado na Ref. 7. Consiste na elaboração e avaliação de um gráfico, onde se destacam a mediana, os quartis e os pontos mais afastados.
Usa-se a mediana (em vez da média) e a distância interquartílica (em vez do desvio padrão), para evitar a influência dos pontos extremos. Isto tornaria o método "robusto", não sendo necessário, para a análise, excluir os pontos discrepantes. São estabelecidos dois limites:

Ponto discrepante: $x_d = m_d \pm 1,5$ deq
Ponto muito discrepante: $x_d = m_d \pm 3,0$ deq

Onde m_d é o valor da mediana e deq a distância interquartílica. Pela curva normal, os quartis distam da média em 0,674 σ. Portanto, a distância interquartílica equivale a
$2 \times 0,674\ \sigma = 1,346\ \sigma$. Os limites são, então:

$x_d = \bar{x} \pm 1,5 \times 1,346\ \sigma = \bar{x} \pm 2,02\ \sigma$ (discrepante)
$x_d = \bar{x} \pm 3,0 \times 1,346\ \sigma = \bar{x} \pm 4,04\ \sigma$ (muito discrepante)

Estes limites são estabelecidos sem qualquer referência ao tamanho da amostra.
Conforme enfatizado no presente trabalho, a amplitude de uma amostra normal aumenta com o tamanho da amostra. A amplitude aumenta indefinidamente. Por exemplo, o valor de 2,02σ, considerado "discrepante" no método box&whisker, já é atingido em uma amostra normal de 17 pontos. A partir deste tamanho, todas as amostras normais conteriam pontos "discrepantes".
Evidentemente, é incorreto estabelecer como limite um determinado afastamento da média, sem considerar o tamanho da amostra. O método box&whisker apenas indica que um determinado valor está relativamente longe da média. Isso não é suficiente;
o ponto só deverá ser considerado discrepante se o seu afastamento não for justificado pelo tamanho da amostra.
O método também não é sempre "robusto" como pretende ser. Em amostras pequenas (menos de seis pontos), os quartis são influenciados pelo ponto discrepante e o método dá resultados evidentemente falsos.
Por exemplo, considerando uma amostra hipotética de 4 pontos:
$x_1 = 8$ $x_2 = 10$ $x_3 = 12$ $x_4 = 24$
O valor do ponto x_4, sendo o dobro do ponto x_3, é obviamente discrepante.
Aplicando o método, obtemos:
$m_d = (10 + 12) / 2 = 11$
quartil superior = (12 + 24) / 2 = 18
quartil inferior = (8 + 10) / 2 = 9
deq = 18 - 9 = 9
$x_d = m_d + 1,5$ deq $= 11 + 1,5 \times 9 = 24,5$
Como $x_d > x_4$, o método box&whisker falhou em identificar o ponto x_4 como discrepante.

O método proposto, aplicado ao mesmo exemplo, identifica corretamente o ponto x_4:
média = 13,5
desvio padrão = 7,188
diferença = 24-13,5 = 10,5
z = (24-13,5) / 7,188 = 1,461
da tabela 3, para N=4, obtém-se z_d = 1,391
como z é maior que z_d, o ponto x_4 é discrepante.

Pelo exposto, o método box&whisker não serve para identificar pontos discrepantes. Apenas chama a atenção sobre os pontos muito afastados da média.

ANEXO 7 (da terceira parte)

SIMULAÇÃO DE AMOSTRAGENS REAIS

Foi simulada a obtenção de amostras aleatórias da distribuição normal, com média 10 e desvio padrão 1, usando a função =INV.NORM(ALEATÓRIO();10;1) da planilha Excel 2010. Foram obtidas quatro mil amostras de cada tamanho, sem pontos discrepantes. Calcularam-se as médias sucessivas; a diferença entre as médias foi dividida pelo desvio padrão calculado com a amostra. Para cada tamanho de amostra foram determinados os valores de d/σ correspondentes a 50% (mediana) e a 100% (totalidade) dos pontos. O resultado é mostrado na tabela a seguir.

N	100%	50%
1	0,7071	0,7071
2	0,5603	0,4609
3	0,463	0,2725
4	0,3905	0,1976
5	0,3339	0,1548
6	0,2916	0,1311
7	0,2583	0,1055
8	0,234	0,0936
9	0,2137	0,0817
14	0,1494	0,0529
19	0,1167	0,0372
24	0,0961	0,0286
29	0,0805	0,0246
39	0,0632	0,0183
49	0,052	0,0142
99	0,0274	0,007
199	0,0147	0,0035

No gráfico 8 abaixo estão indicados os valores da mediana (bolinhas azul claro) e da totalidade dos pontos (bolinhas brancas). Os valores experimentais convergem para o

valor 0,7071 em N=1, porque as estimativas da média e do desvio padrão não são independentes.
O tamanho da amostra conforme a equação [12] também consta neste gráfico como uma linha amarela. Pode-se observar que todos os resultados experimentais ficam abaixo desta linha. Assim, a simulação confirma que os tamanhos indicados são adequados para os casos reais.

a)determinação do fator de correção h
Para considerar o efeito do tamanho da amostra sobre z' na equação [11], é necessário ajustar a equação aos dados experimentais. O fator de correção será determinado considerando que os três últimos valores da tabela acima (coluna 100%) são um pouco mais altos que os calculados com a equação [11]. Selecionando o valor experimental de d/σ correspondente a N=99, com z' = 2,58, o fator de correção da equação é

$$h = \text{experimental} / \text{teórico} = 0{,}0274 / (2{,}58 / (99 + 1)) = 1{,}062$$

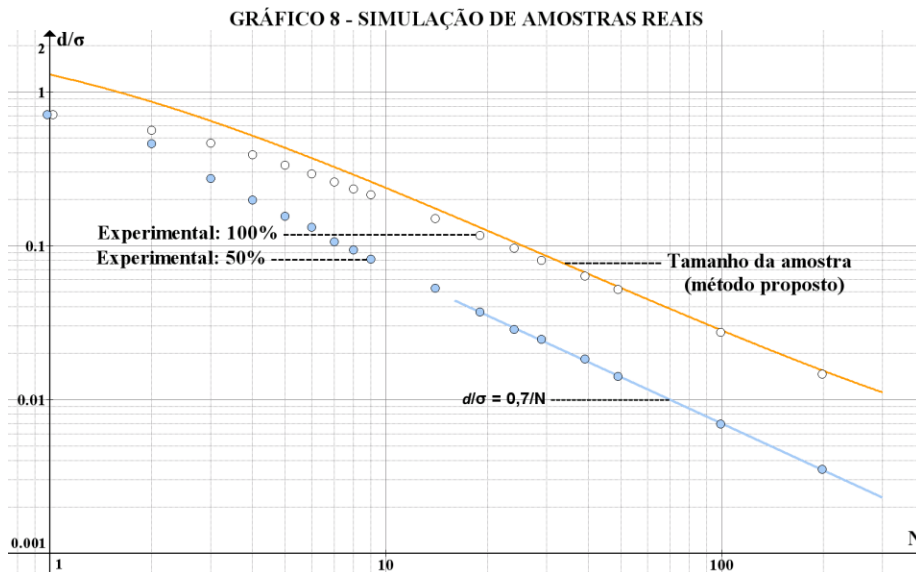
colocando em função linear de N:

$$h = 1 + 0{,}062 \times N / 99 = 1 + 0{,}00063\ N \approx 1 + 0{,}001\ N$$

Com este fator de correção, os valores da equação [12] e tabela 4 ficam acima de todos os valores experimentais. Portanto, o nível de confiança do método é superior a 99%.

b) relação entre os valores de d/σ e o tamanho da amostra
Os valores experimentais medianos são representativos da evolução média esperada dos valores de d/σ com o tamanho da amostra. Para os valores maiores de N, foi ajustada uma reta (em azul, no gráfico 8), correspondente à equação empírica

$$d/\sigma = 0{,}7 / N \qquad \text{equação [13]}$$

Em média, os valores de d/σ variam inversamente proporcionais ao tamanho da amostra. Esta equação é utilizada no anexo 8, para estimar o limite máximo do tamanho da amostra.

d/σ: diferença admissível / desvio padrão estimado
N: número de pontos da amostra
Obs: gráfico log-log

ANEXO 8 (da terceira parte)

LIMITES MÁXIMOS PARA O TAMANHO DA AMOSTRA

Conforme visto na terceira parte, a relação d/σ e, conseqüentemente, a contribuição de cada ponto adicional para a precisão do resultado, diminui à medida que o tamanho da amostra aumenta. Quando se torna desprezível ou nula, não há vantagem em aumentar o número de pontos. Foi atingido um limite máximo razoável para o tamanho da amostra.

a) máximo quando os custos são significativos
Quando os custos de amostragem são significativos, somente devem adicionados pontos à amostra enquanto a contribuição dos mesmos para a melhoria da precisão for importante.
No início, o efeito de N sobre a precisão é muito acentuado, mas diminui rapidamente à medida que N aumenta. Conforme a tabela 4 do item 3.4, a diferença (d/σ) varia entre 1,29 (para N=1) e 0,015 (para N=200). Portanto, a maior redução possível na diferença é 1,29 - 0,015 = 1,275. Verifica-se que uma grande parte desta redução já é alcançada com amostras relativamente pequenas. Com 30 pontos a diferença é 0,086; logo,a redução é 1,29 - 0,086 = 1,204, ou cerca de 94% da redução possível (neste raciocínio, estamos admitindo que o tamanho máximo da amostra é 200).
Enquanto com 30 pontos se atingiu 94% da redução possível, para os 6% restantes seriam necessários mais 170 pontos. Vê-se que a melhoria na precisão dificilmente justificaria os custos adicionais. Portanto, quando os custos de amostragem são significativos, um limite razoável para o tamanho da amostra é de 30 pontos.

b) máximo quando os custos não são significativos
Veremos a seguir que, mesmo quando os custos são baixos, há também um limite, a partir do qual a melhoria na precisão se torna desprezível.

– Exatidão de medições
Uma informação importante, relativa à exatidão que pode ser alcançada em medições, encontramos em Vuolo (Ref. 3, pág. 68). Como regra, quando se expressa a média de uma série de medições, o resultado deve ser acrescido da incerteza padrão (desvio padrão da média), que deve ser indicada com 2 algarismos significativos, se o primeiro algarismo for 1 ou 2. Mais de 2 algarismos significativos não tem utilidade prática, porque raramente se consegue uma exatidão maior.
Se o primeiro algarismo for 2, uma unidade no segundo algarismo representa 5% do valor da incerteza padrão. Portanto, o limite do erro no desvio padrão da média é 5%.
Esse autor também enfatiza que, ao expressar um resultado na forma $\bar{x}$ = média ± incerteza, os algarismos significativos devem ser consistentes, prevalecendo a precisão do valor menos exato. Por exemplo, para expressar corretamente um resultado tal como: $\bar{x}$ = 10,02 ± 1,0 o algarismo 2 não é consistente e deve ser ignorado.

– Conseqüência para o tamanho da amostra
Como as informações acima podem ser usadas para estabelecer um limite?
Sabemos que a diferença entre as médias diminui com o tamanho da amostra, aproximadamente proporcional a 1/N. Atingido o ponto em que esta diferença é menor que 5% do desvio padrão da média, deve prevalecer a precisão deste último.
O erro no desvio padrão da média é igual a σ/√N. Então, a partir deste ponto, as diferenças diminuem proporcionalmente a 1/√N. Este ponto pode ser determinado como segue.
O desvio padrão da média (com σ=1) é $\sigma_m = 1 / \sqrt{N}$
Pelas simulações realizadas (equação [13], anexo 7), a mediana das diferenças obtidas corresponde à relação empírica (σ = 1):
d = 0,7 / N
Se a diferença é igual a 5% do desvio padrão da média, pode-se escrever
0,7 / N = 0,05 / (N^0,5)
Donde se obtém
N = 196
A partir deste tamanho, as reduções nas diferenças, que antes eram proporcionais a 1/N, passam a ser proporcionais a 1 / √N. O efeito do tamanho da amostra é muito menor e a melhoria na precisão se torna desprezível. Conclui-se que, mesmo se os custos de amostragem forem baixos, não há interesse em obter amostras com mais de aproximadamente 200 pontos.

c) máximo absoluto
A existência de um máximo absoluto também é apontada por Vuolo (Ref. 3, pág. 110). A incerteza padrão (desvio padrão da média) apresenta dois componentes: a incerteza estatística e a incerteza sistemática residual. A incerteza sistemática residual resulta do fato que não é possível eliminar completamente os erros sistemáticos de uma medição.
A incerteza estatística pode ser diminuída aumentando-se o número de medições.

A incerteza sistemática não varia. Assim, para um número muito grande de pontos, a incerteza estatística fica menor que a incerteza sistemática residual e esta última estabelece um limite final para a exatidão do resultado. Foi atingido o ponto em que é totalmente inútil aumentar o tamanho da amostra.
A equação [12] do método proposto é coerente com este fato. Para valores muito grandes de N, o valor calculado de d/σ tende a ficar constante. Já no método da literatura, conforme a equação [8], o erro poderia ser reduzido infinitamente.

REFERÊNCIAS

1) Ullmann's "Encyklopädie der Technischen Chemie", vol. 2/1.
Urban&Schwarzenberg. München, 1961.

2) Stevenson, W. J. "Estatística Aplicada à Administração"
HARBRA. São Paulo, 1981

3) Vuolo, J. H. "Fundamentos da Teoria de Erros"
Edgar Blücher, São Paulo, 1996

4) Pillar, V.D. "Suficiência Amostral"
Departamento de Ecologia da Universidade Federal do Rio Grande do Sul
Porto Alegre, 1999

5) Moroney, M. J. "Facts from Figures".
Penguin Books, Harmondsworth, Middlesex, 1951

6) Grubbs, E. F. "Procedures for Detecting Outlying Observations in Samples"
Technometrics, Vol. 11, No. 1 (Feb. 1969). American Statistical Association.

7) Petrobrás SEREC/CEN-SUD, "Curso de Estatística Básica", Parte 1.
Rio de Janeiro, 1992

No presente trabalho foram usados a planilha Microsoft Excel 2010 e o desenhador de gráficos Graph, versão 4.4.2, http://www.padowan.dk/